基于聚类和流量传播图的 P2P 流量识别方法 *

苏阳阳 [a]，孙冬璞 [a]，李丹丹 [a, b]，孙广路 [a, b]

(哈尔滨理工大学 a. 计算机科学与技术学院; b. 信息安全与智能技术研究中心, 哈尔滨 150080)

摘 要：为有效监管网络，快速精确识别 P2P 流量，通过分析 P2P 网络流量中节点与节点、节点与链路之间的交互和行为特征，将聚类方法与流量传播图方法相结合，提出了一种基于网络行为特征的 P2P 流量识别方法。该方法首先通过采集网络流的包级和流级统计特征对不同种类的网络应用的流量进行聚类，然后利用流量传播图对 P2P 流量进行识别。实验结果表明，提出的方法在骨干网络数据上能够有效识别 P2P 网络应用流量，F1-measure 达到 95%以上。

关键词：P2P 流量识别；流量行为特征；流量传播图；基于密度带噪声的空间聚类算法

中图分类号：TP393　　doi: 10.3969/j.issn.1001-3695.2018.06.0328

P2P traffic identification method based on clustering And Traffic Dispersion Graph

Su Yangyang[a], Sun Dongpu[a], Li Dandan[a, b], Sun Guanglu[a, b],

(*a. School of Computer Science & Technology*; *b. Research Center of Information Security & Intelligent Technology*, *Harbin University of Science & Technology*, *Harbin* 150080, *China*)

Abstract: In order to effectively supervise the network, quickly and accurately identify the peer-to-peer flow, by analyzing the interaction and behavior characteristics between nodes and nodes, nodes and links in Peer-to-peer network traffic, a method of Peer-to-peer traffic recognition based on network behavior features is proposed by combining clustering method with flow propagation graph method. Firstly, the flow rate of different kinds of network is collected by collecting packet level and flow level statistic feature of network flow, and then the Peer-to-peer flow is identified by using traffic graph. The experimental results show that the proposed method can effectively identify Peer-to-peer network application traffic in backbone network data, and the f1-measure reaches over 95%.

Key words: P2P traffic identification; traffic behavior characteristics; traffic dispersion graph; density-based spatial clustering of applications with noise

0 引言

对等网络(peer-to-peer,P2P)是一种无须经过中间实体的网络模型。近年来，伴随着计算机网络的迅猛发展，许多网络应用采用 P2P 技术原理来实现其服务。因此，P2P 协议已经广泛应用于即时通信、视频分享、文件共享、在线直播、游戏等领域。通过 Cisco(思科)公司的年度流量统计报告表明，虽然 P2P 流量占据全球总体网络带宽的比率有下降趋势，但仍可以达到带宽总量的 40%[1]。由于 P2P 应用采用多网络连接模式，进而可以保证其数据传输效率，但其大量占用网络带宽，易引发网络拥塞等问题。因此，能够在全网流量中精确的辨识出 P2P 流量并对其进行有效地监管具有非凡的意义。

现在已有的 P2P 流量识别方法主要依赖对 P2P 网络和流量自身的特点进行研究与分析，发现其特有的静态与动态特征，将 P2P 流量与网络中其他流量有效区分，以帮助网络管理者和服务商提升不同网络业务的服务质量。针对 P2P 流量识别方法主要包括基于端口的识别方法、基于载荷特征的识别方法、基于流统计特征和机器学习的识别方法以及基于网络节点关系和主机行为的识别方法[2]。这些方法分别根据不同的角度对 P2P 流量进行了分析与识别，各有优缺点。由于现有的 P2P 应用大多使用动态端口和加密方式进行传输，使得基于端口和载荷特征的方法无法进行有效识别[3,4]。而基于流统计特征和机器学习的识别方法虽然不是单纯的依赖于端口和载荷，但是由于流统计特征在不同网络环境下的取值范围的不稳定性，会使训练数据与测试数据间差异较大，进而影响有监督机器学习模型的识别效率。而且对于不同网络环境中新出现的协议的适应性较差[5,6]。

收稿日期：2018-06-15；**修回日期：**2018-07-29　**基金项目：**国家自然科学基金资助项目（60903083，61502123）；黑龙江省新世纪人才项目（1155-ncet-008）；黑龙江省博士后科研启动基金资助项目

作者简介：苏阳阳（1992-），男，硕士研究生，主要研究方向为信息安全、数据挖掘（415828587@qq.com）；孙冬璞（1979-），女，副教授，博士，主要研究方向为数据挖掘、数据库技术；李丹丹（1989-），女，硕士研究生，主要研究方向为信息安全、机器学习；孙广路（1979-），男，教授，博士，主要研究方向为信息安全、人工智能.

对于基于网络节点关系和主机行为的方法虽然是可以识别新的协议，但是受限于网络拓扑环境的变化，难以应用到高速骨干网中。Iliofotou 等人[7]提出了流量传播图（traffic dispersion graphs，TDG）的概念，将节点之间的通信关系转换为有向图，挖掘深层次的网络交互行为，量化有向图中的入度、出度、网络直径、最大连接组件等特征，并利用这些特征来识别有向图中通信链路的应用类型。可是网络中的通信链路不完全都是能够相互连通的，并且即使相互连通的链路也不一定在相同时间内都是属于同一种网络应用。因此，如果单独使用 TDG 方法进行不同流量的识别，可能对于特征属性不够明显的小流量会被错误识别，甚至会出现不能被识别的问题。

因此，本文提出了 CTDG（clustering and traffic dispersion graph based method），一种改进的聚类与 TDG 图模型相结合的 P2P 流量识别方法。CTDG 方法首先通过无监督机器学习模型将采集到的网络流量中具有相似统计特征的流聚类为若干潜在可能被识别的类；然后利用 TDG 图中定义的度量量化网络流量之间的交互行为为特征，并在此基础上进行 P2P 流量识别。实验结果表明，本文提出的方法在高速骨干网中对 P2P 流量识别效果明显，其准确率能够达到 95%以上。

1　相关工作和存在问题

在现有的 P2P 的流量识别方法中，基于流统计特征和机器学习的识别方法不依赖于应用层载荷内容，而是基于网络层和传输层分析并提取流量统计特征，结合带标记的流量数据集，在有监督的机器学习模型中进行模型训练，最终识别各种应用产生的流量。该方法通常使用数据包级（packet level）特征和数据流级（flow level）特征。其中，数据包级特征主要包括端口号、数据包平均到达时间、以太网数据包最大字节数以及数据包间最大时间间隔等；数据流级特征主要包括单个数据流的持续时间、长度以及与其他流的间隔时间等。徐洪平等人[8]通过构建动态混合识别策略结合 SVM 和投票机制对流量进行识别。Roughan 等人[9]提出了基于上述统计特征的最近邻和线性判别分析方法。Liu 等人[10]提出了 26 种 P2P 流的统计特征，并利用支持向量机模型区分四种 P2P 流量，得到较好的识别效果；但对流数较少的应用类别，难以进行有效的识别。孙知信等人[11]提出了一种基于流特性描述的 P2P 游戏流量识别方法，通过对标注的关键性流量进行数据包分布情况的分析，利用隶属度函数作为评语集，最后利用模糊评判的准则判定流量的网络应用来识别 P2P 流量；但该方法着重依赖于载荷数据，对识别特征不明显和加密的 P2P 网络应用的适应性较差。陈阳[12]提出了基于 SVM 的 P2P 流量早期识别研究，利用数据流早期数据包进行特征选择和识别。戴磊等人[13]则是通过使用主动学习技术提取出少量具有高质量的样本，利用支持向量机模型建模进行 P2P 流量识别。但如何将其应用于实际的复杂网络环境，还需根据具体问题具体分析。

对于基于网络节点关系和主机行为的网络流量分类方法重点是依照主机在网络中所承担的作用和各个主机之间的连接方式以及某些网络中群体行为等方面来考虑。Karagiannis 等人[14]率先提出使用 P2P 网络中对等体间连接的模式来对网络中的 P2P 流量进行识别[12]，随后他又提出了基于主机行为模式的网络流量分类方法（BLINC）。BLINC 方法将主机的行为模式分为社会层、功能层、应用层，并通过提取这些行为模式来识别网络流量。该方法虽然提高了 P2P 流量的匹配度，但过度依赖于端口和 IP 间的关系。胡斌[15]提出了基于混合行为特征与 Spark 大数据并行框架相结合的流量识别方法。Constantinou 等人[16]根据记录中每个节点与其他节点建立连接的实际情况来获取 P2P 网络的连接拓扑图，通过计算它的网络直径并与其他类型网络的拓扑图比较发现，P2P 网络拓扑图的网络直径会更大，从而致使该方法对需求数据的处理及度量计算体系的要求会很高，难以达到方便易用的结果。鲁文斌等人[17]针对 P2P 网络的分布式特点，利用单位时间内结点与目的子网间的连接数，以及连接数与有效连接数的比值等特性，提出一种基于节点连接特性的 P2P 节点识别算法。该算法处理时间虽然比深度报文检测的时间要短，但更多的依赖传输层的特征对 P2P 流量进行识别。

针对上述不足，本文提出了一种改进的聚类与 TDG 图模型相结合的 P2P 流量识别方法（clustering and traffic dispersion graph based method，CTDG）。该方法有以下优点：a）不需要使用载荷内容，能够识别加密的 P2P 流量；b）挖掘深层的 P2P 网络交互行为以此区别于其他应用网络的图特征，并据此有效的识别 P2P 应用；c）对于网络中新出现的应用，具有很好的适用性，不需要训练和配置复杂的模型参数。

2　网络流量的统计特征提取

本文采用网络中常用的五元组信息{源*IP*、目的*IP*、源端口、目的端口、传输层协议}来定义网络流，使用一定时间段内的双向流作为基本单元。以首个数据包的发送端作为源端、接收端作为目的端来定义 TCP 流的方向。以相同五元组流量中第一个数据包的发送端作为源端、接收端作为目的端来定义 UDP 流的方向。本文分别提取了网络流的统计特征、网络节点关系和主机行为特征。

不同应用层协议产生的网络流在数据包级特征和数据流级特征上会有比较明显的差异[18]。本文提取了数据流大小、会话的持续时间、流中每个数据包到达的时间、双向的数据包数目、包到达的时间间隔（均值、方差）以及通信双方在 idle 上花费的时间等 60 种网络流统计特征，并使用信息增益算法[19]抽取最相关的特征作为聚类属性，如表 1 所示。

表 1　流统计特征

数据包级的特征	数据流级的特征
前 6 个数据包的字节长度	流大小
最大，最小数据包长度	流的持续时间
数据包长度平均值，方差	流到达间隔时间

3　流量传播图（TDG）

本文用一个有向图 $G(V,E)$ 定义全部网络节点之间的 TDG 图，节点集 V 表示网络中的节点集，图中的边 $edge(u,v) \in E$ 代表主机 u 发向主机 v 的网络流。

P2P 网络中每个节点能够决定自身通信行为，具有独立性；但是节点间通过链路通信进行协作以获取信息和计算资源，又具有相互依赖性。P2P 网络节点的 TDG 图（图 1）具备以下特点：

a）节点平均度非常高。这是因为大量 P2P 节点之间通过相互连接来实现数据共享和内容查询。

b）同时拥有(出、入)度的节点在网络整体中所占的比重较大。这是因为网络中大量 P2P 节点同时拥有服务器和客户端双重身份的特性所决定的。

c）部分 P2P 网络的网络直径会很大。这是因为 BitTorrent 等 P2P 应用具有分散式网络拓扑结构。

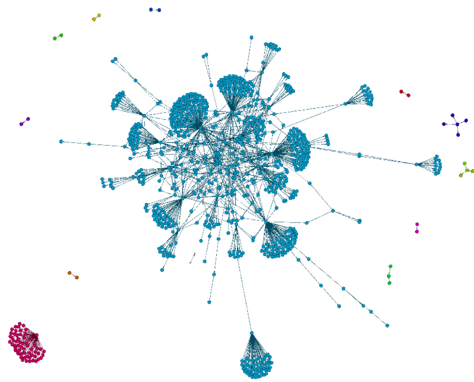


图 1　P2P 网络的 TDG 图

基于 TDG 图定义和 P2P 网络的图模型特征分析，本文构建的行为特征包括：

a）同时具有入度和出度的节点占所有节点数的百分比，用 IO 表示。

b）所有的节点之间的最短距离中路径最长的两个节点之间的距离，即网络直径。

c）节点平均度：$2|E|/|V|$。

4　DBSCAN 聚类算法

基于密度带噪声的空间聚类方法（density-based spatial clustering of application with noise，DBSCAN）是一种典型的密度聚类算法。算法的核心思想是通过密度可达关系导出的最大密度相连的样本集合，作为最终要得到的一个类别。其通过一组邻域来描述样本集的紧密程度，使用参数 eps 和 minPTS 用来描述邻域的样本分布紧密程度，参数 eps 描述了某一样本的邻域距离阈值，参数 minPTS 描述了某一样本的距离为 eps 的邻域中样本个数的阈值。算法基本流程如下：

a)DBSCAN 聚类算法需要获得样本中全部的核心对象即对每一个样本根据距离度量方式，获得满足 eps 邻域距离和大于 minPTS 样本个数的样本作为核心对象成员。

b)在核心对象集合中，随机选择一个对象，初始化当前簇核心对象队列、类别序号、当前簇样本集合以及未访问的样本集合。通过迭代选取当前簇核心对象队列中的每一个对象，根据邻域距离阈值 eps 找出此对象邻域子样本集合，用于更新当前簇的样本集合和未访问的样本集合，同时将邻域子样本集合中是核心对象的样本加入到当前簇的核心对象队列中。

c)如果当前簇的核心对象队列不再增加，则当前簇生成为一个新类别 C_k，加入到簇划分集合 $C=\{C_1,C_2,\ldots,C_K\}$中。直到核心对象中的每一个成员被划分到簇划分集合中，聚类结束。因为 DBSCAN 算法能够识别不同形状的簇，而且对噪声点具有较强的鲁棒性[20]，所以本文利用它来进行流量识别前的网络流分流处理。

不同的距离计算方法会对 DBSCAN 算法的聚类效果产生直接的影响，传统的 DBSCAN 算法使用欧氏距离作为距离的度量方式，欧氏距离的度量方式更多地关注了各个特征值间的绝对距离，往往忽视了样本间的相对距离。P2P 网络数据流间相对距离的比较，更能准确地刻画样本间存在的相对联系。所以本文提出 DBSCAN 算法利用卡方距离度量各个样本间的相对距离。卡方距离公式为

$$dx^2(x,y)=\sum_n \frac{(x_n-y_n)^2}{(x_n+y_n)}$$

卡方距离是根据卡方统计量提出的，已经被广泛应用于实际距离度量问题中，并且取得了相当好的效果[21]。

5　CTDG 流量识别方法

CTDG 方法结合了改进后的聚类和 TDG 图的关系挖掘方法，使用网络的流量统计特征和行为特征来实现 P2P 流量的有效识别。CTDG 方法的识别流程如图 2 所示。其可以分为以下四个步骤。

a)过滤。

因为基于端口和载荷的方法识别某些非加密的传统应用效果较好，所以本文应用基于端口和载荷的识别方法将 Web、DNS 和 SMTP 等可以识别的应用过滤掉。这不仅减少了其他背景流量的干扰，也能降低后续步骤的时间和空间复杂度。

b)分流。

利用表 1 列出的统计特征，使用 DBSCAN 作聚类，将统计特征相近的网络流聚成簇。算法选取欧几里得度量计算特征空间中的相似度。设 F 代表网络流数据集，$f_i \in F$ 代表其中的每一个网络流，算法的详细步骤如下：

（a）对数据集中的每一个未处理的数据流 $f_i \in F$，划定其扫描半径（eps），检测 eps 涵盖范围内的数据流。若其个数大于最小流数阈值（minPTS），则创建新簇 Y，将这些数据流加入 Y 簇中。

（b）对 Y 簇中每一个的数据流 y_j，检测其 eps 涵盖范围内的数据流。若其个数大于等于 minPTS，则将其中没有包含在任何簇的数据流聚入 Y 簇中。

（c）重复执行步骤（b），直至没有新网络流聚入簇 Y。

（d）根据识别的结果，重复执行步骤（a）~（c），直到所有的网络流都被处理。

c)合并相似簇。

将 IP 相似性定义为：两个簇中出现相同的 IP 的个数和两个簇中 IP 的总数的比值。

如果当 IP 相似性难以满足预先设定的阈值时，则结束合并。

理想的聚类结果为，相同应用产生的数据流应被聚到同一个簇中，而且一个簇中只包含一种应用流量。但是在实际聚类结果中发现，相同应用也会产生了多个簇。通过分析发现，P2P 协议具有多种交互模式，在查询过程中一般使用 UDP 协议进行通信，在文件传输过程中使用 TCP 协议进行通信，这两种通信模式在包级和流级统计特征上有很大差别。由于相同应用产生的不同簇对应的 TDG 会有大量的共同节点，本文将 IP 相似性作为簇的合并条件。

d)利用合并结束后的每组流创建 TDG，并利用其度量指标进行分类。TDG 的度量指标为：利用本文构建的 TDG 图的行为特征，同时具有入度和出度的节点占所有节点数的百分比，网络直径大小限制以及节点平均度作为 TDG 分类指标。

将上述过程中得到的不同簇创建为 TDG 并计算它们的度量值。如果度量值满足设定的阈值，则判断该 TDG 符合 P2P 模式，并将其中的每个流都标记为 P2P 应用。

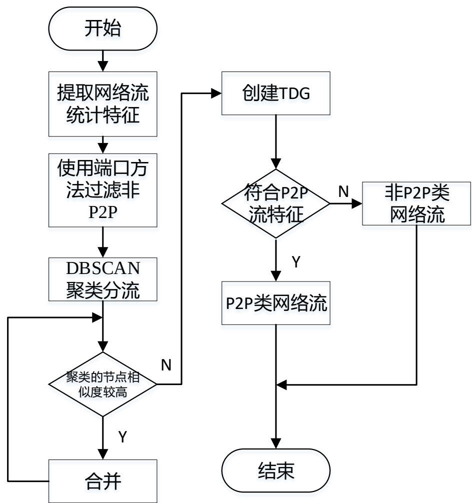


图 2　CTDG 方法流程

6　实验结果和分析

6.1　数据集

本文应用 2017 年不同时间采集于中国某骨干网络的流量作为实验数据集。表 2 进行了详细的实验数据描述。本方法中使用 CoralReef 来处理网络流量。CoralReef 是一种用于被动分析互联网流量的软件套件。将其设定 64 s 为流超时值，并使用基于载荷特征匹配的方法来标注数据集。

表 2　流量数据集信息

数据集	Backbone1	Backbone2
流量持续时间	5 min	30 min
数据包数目	16 000 000	126 000 000
数据包字节数	10 GB	80 GB
流数	2 000 000	19 000 000

经过手工标注和分析，实验数据集主要包括 DNS、Web、P2P、Streaming、Games、Network-operation、MAIL/NEWS 等网络应用协议类型，还有部分载荷分析方法难以识别的应用。在实验过程中，本文删除了难以识别的流以及没有载荷的流。图 3 描述了两份网络流量数据集中应用类型的分布情况。

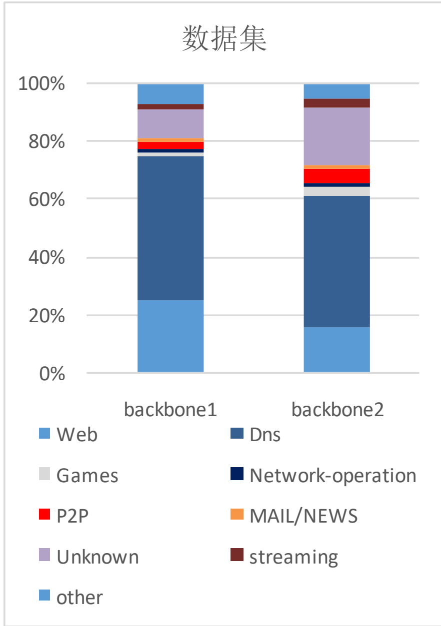


图 3　实验数据中应用协议类型分布

6.2　评价方法

为了准确地评价本文提出的方法，采用准确率（precision）/召回率（recall）和综合评价指标（F1-measure）评价结果。各指标定义如下：

准确率（precision）为

[illegible]　(1)

召回率（recall）为

[illegible]　(2)

综合评价指标（F1-measure）为

[illegible]　(3)

其中：TP（true positives）代表被正确分类为 P2P 的样本数目；FP（false positives）代表将非 P2P 样本被错误识别为 P2P 的数

目；FN（false negatives）则代表将 P2P 样本错误识别为非 P2P 的数目。

6.3　实验结果及分析

首先测试 DBSCAN 算法将属于相同应用的流聚成簇的效果。根据基准方法对样本流的标记类型，选择每个簇中含有标记最多的应用类型来标记这个簇，即簇中的所有流都被标记为该应用类型。DBSCAN 算法经过调节参数 eps 和 minPTS 两个参数来对簇的最终数目和聚类结果进行调整。其中，minPTS 越小会产生越多数量的簇。确定最小 minPTS 后，随着 eps 的增大，分类性能也会不断提升。但是当 eps 过大时，分类性能又会明显减弱。由图 4 所示，eps 在 0.02~0.04，minPTS=4 时算法效果最好，聚类得到的簇标记的准确率达到 90%以上。

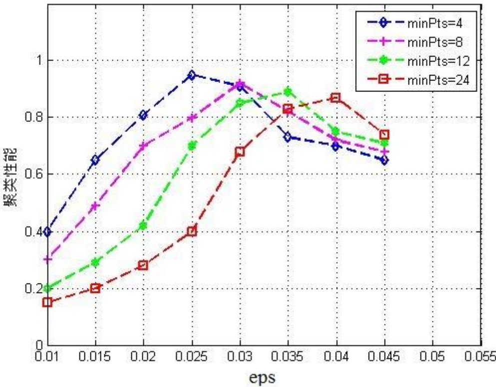


图 4　不同 eps 和 minPTS 参数下的聚类结果

CTDG 方法中，步骤 c)是合并可能运行同种应用的多个簇，合并的效果取决于节点相似度阈值的设定。阈值设定过大，使得多个簇很难合并到一起，导致相同应用的网络流分布在不同的簇中，不利于全面地分析相同类型网络流的行为模式；阈值设定过小，使得不同应用的簇被错误地合并，降低算法的整体识别准确率。

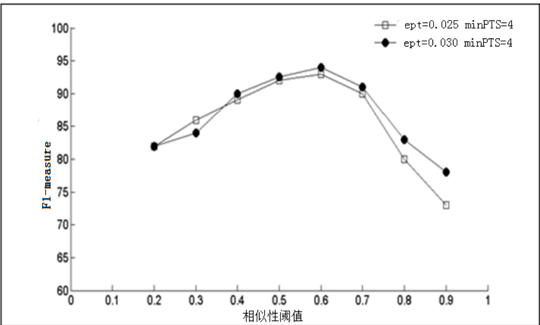


图 5　相似性阈值的选取对识别结果的影响

由图 5 所示，如果把节点相似度阈值设置在 0.4~0.7 时，可以发现 CTDG 分类方法的准确率可以超过 90%，表示分类效果较好。在聚类参数 minPTS=4、ept=0.025、节点相似度阈值设为 0.6 时，CTDG 方法能够达到 93%的召回率和 96%的准确率；在 ept=0.03 时，CTDG 方法也能够达到 90%以上的召回率和准确率。但是在实验过程中也发现，参数选择不当会导致 CTDG 方法的分类性能会大幅度下降。

相对于本文提出的 CTDG 方法，BLINC 方法根据主机在传输层的连接模式（如端口和 IP 的关系等）来标记每个主机的所有流。本文中使用 BLINC 方法对现有数据集作分类，BLINC 能达到 84%的准确率和 89%的召回率。此外，BLINC 对于 BitTorrent 等部分 P2P 应用的识别率比较低，只能达到 25%，而 CTDG 方法的检测率却能达到 90%，如表 3 所示。由于 CTDB 方法引入了聚类过程，并利用了更多的统计特征作为聚类的量度，使得建立 TDG 的效果更好，有效地提升了最终方法的识别性能。相比陈阳提出的利用 SVM_PF 进行 P2P 流量识别，该方法根据双向流的早期多个数据包作为特征选择依据，虽然可以降低特征提取的复杂度，但由于缺少关联性更强的特征，召回率普遍在 85%以下，如图 6、7 所示。

表 3　CTDG,SVM_PF,BLINC 方法的性能对比/%

方法	准确率	召回率	F1-measure
BLINC	84.5	89.7	87.0
SVM_PF	93.2	83.1	88.3
CTDG	96.8	93.6	95.1

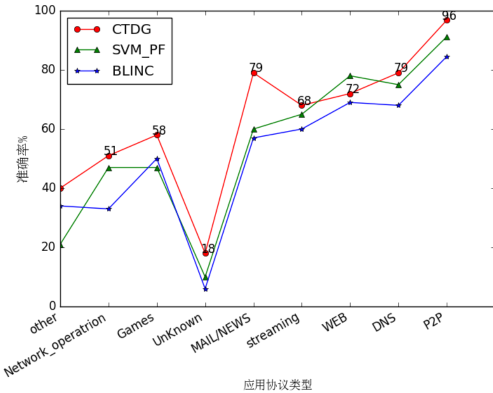


图 6　准确率比较

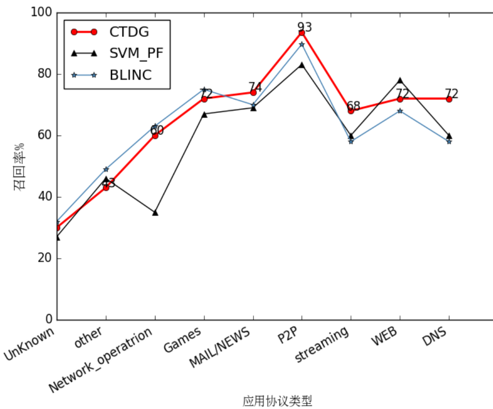


图 7　召回率比较

7　结束语

本文针对 P2P 流量的网络行为特性，应用基于网络流的包级和流级统计特征的聚类方法，并结合 TDG 图特征，提出了将网络流和主机行为特征与 TDG 相结合的 CTDG 方法用于 P2P 流量识别，对 P2P 流量的识别更为精准有效。通过实验表明，该方法较 BLINC 和 SVM_PF 方法在准确率、召回率和 F1-measure 方面均有不错的提升；同时本文的贡献在于为解决

传统 P2P 流量识别的问题提供了新的研究思路。

参考文献：

[1] Cisco Systems. Cisco visual networking index: forecast and methodology, 2010-2015 [R]. 2011.

[2] Kim H, Claffy K, Fomenkov M, *et al.* Internet traffic classification demystified: myths, caveats, and the best practices [C]// Proc of ACM CoNEXT Conference. [S. l.] : ACM Press, 2008: 1-12.

[3] 孙恩博. 高速网络环境下的 P2P 僵尸网络检测方法研究 [D]. 成都: 电子科技大学, 2016. (Sun Boen. Research on detection method of Peer-to-peer botnet in high-speed network environment [D]. Chengdu: University of Electronic Science and Technology, 2016.)

[4] 牛伟纳, 张小松, 孙恩博, 等. 基于流相似性的两阶段 P2P 僵尸网络检测方法 [J]. 电子科技大学学报, 2017, 46 (6): 902-906, 948. (Niu Weiba, Zhang Xiaosong, Sun Boen, *et al.* Two-stage peer-to-peer zombie network detection method based on flow similarity [J]. University of Electronic Science and Technology, 2017, 46 (6): 902-906, 948.)

[5] 王春枝, 杜远丽, 叶志伟. 基于最优 ABC-SVM 算法的 P2P 流量识别 [J]. 计算机应用研究, 2018, 35 (2): 582-585. (Wang Chunzhi, Du Yuanli, Ye Zhiwei. Peer-to-peer traffic recognition based on optimal ABC-SVM algorithm [J]. Application Research of Computers, 2018, 35 (2): 582-585)

[6] Dainotti A, Pescape A, Claffy K. Issues and future directions in traffic classification [J]. IEEE Network, 2012, 26 (1): 35-40.

[7] Iliofotou M, Pappu P, Faloutsos M, etal. Network monitoring using traffic dispersion graphs [C]// Proc of ACM SIGCOMM conference on Internet measurement. [S. l.] : ACM Press, 2007: 315-320.

[8] 徐洪平, 刘洋, 易航, 等. 运载火箭测发网络异常流量识别技术 [J]. 清华大学学报: 自然科学版, 2018, 58 (1): 20-26, 34. (Xu Hongping, Liu Yang, Yi Hang, *et al.* Anomaly flow identification technology forlaunch vehicle detection network [J]. Journal of Tsinghua University: Natural Science, 2018, 58 (1): 20-26, 34.)

[9] Roughan M, Sen S, Spatscheck O, *et al.* Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classify-cation [C]// Proc of ACM SIGCOMM conference on Internet measurement. [S. l.] : ACM Press, 2004: 135-148.

[10] Liu H, Feng W, Huang Y, *et al.* A peer-to-peer traffic identification method using machine learning [C]// Proc of International Conference on Networking, Architecture and Storage. 2007.

[11] 孙知信, 宫婧. 一种基于流特性描述的 P2P 流量模糊识别方法 [J]. 计算机学报, 2008, 31 (7): 1252-1260. (Sun Zhixin, Gong Jing. A method of fuzzy recognition of peer-to-peer traffic based on flowcharacteristic description [J]. Chinese Journal of Computers, 2008, 31 (7): 1252-1260.)

[12] 陈阳. 基于 SVM 的 P2P 流量早期识别研究 [D]. 保定: 河北大学, 2017. (Chen Yang. Research on the early recognition of Peer-to-peer traffic based on SVM [D]. Baoding: Hebei University, 2017)

[13] 戴磊, 云晓春, 张永铮, 等. 一种基于 TCM 主动学习的 P2P 流识别技术 [J]. 高技术通讯, 2010 (7): 23-29. (Dai Lei, YunXiaochun, Zhang Yongzheng, *et al.* A peer-to-peer flow recognition technique based on TCM active learning [J]. Hi-Tech Newsletter, 2010 (7): 23-29.)

[14] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark [J]. ACM SIGCOMM Computer Communication Review, 2005, 35 (4): 229-240.

[15] 胡斌. 基于混合行为特征的流量识别技术研究与应用 [D]. 北京: 北京邮电大学, 2017. (Hu Bin. Research and application of traffic recognition technology based onmixed behavior characteristics [D]. Beijing: Beijing University of Posts and Telecommunications, 2017)

[16] Constantinou F, Mavrommatis P. Identifying known and unknown peer-to-peer traffic [C]// Proc of IEEE International Symposium on Network Computing and Applications. [S. l.] : IEEE Press, 2006: 93-102.

[17] 鲁文斌, 杨家海, 刘洪波. 基于节点连接模式的 P2P 节点识别算法 [J]. 清华大学学报: 自然科学版, 2009, 49 (7): 1045-1049. (Lu Wenbin, Yang Jiahai, Liu Hongbo. Peer-to-peer node recognition algorithm based on node connection mode [J]. Journal of Tsinghua University: Natural Science, 2009, 49 (7): 1045-1049.)

[18] Nguyen T, Armitage G. A survey of techniques for internet traffic classification using machine learning [J]. IEEE Communications Surveys & Tutorials, 2008, 10 (4): 56-76.

[19] 李玲, 刘华文, 徐晓丹, 等. 基于信息增益的多标签特征选择算法 [J]. 计算机科学, 2015, 42 (7): 52-56. (Li Ling, Liu Huawen, Xu Xiaodan, *et al.* A multi-label feature selection algorithm based on information gain [J]. Computer Science, 2015, 42 (7): 52-56)

[20] Xie G, Iliofotou M, Keralapura R, *et al.* Subflow: towards practical flow-level traffic classification [C]// Proc of IEEE INFOCOM. [S. l.] : IEEE Press, 2012: 2541-2545.

[21] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporaldata [J]. Data & Knowledge Engineering, 2007, 60 (1): 208-221.